



# APPLICATION FOR LETTERS PATENT OF THE UNITED STATES

## SPECIFICATION

To all whom it may concern:

Be It Known, That I, **Ramin C. Nakisa**, of Bucks, GB, have invented certain new and useful improvements in **PREDICTING FUTURE BEHAVIOR OF AN INDIVIDUAL**, of which I declare the following to be a full, clear and exact description:

## **PREDICTING FUTURE BEHAVIOR OF AN INDIVIDUAL**

### **Background of the Invention**

5 This invention relates to a method and a computer program for predicting the future behavior of an individual, and is particularly, although not exclusively, useful in customer relationship management for automatically maintaining a relationship between a business and its customers through predictive modeling and data mining.

Customer relationship management (CRM) applications take many forms such as computer programs for effecting direct marketing campaigns and recommendation engines. 10 Although a large volume of information is captured in many industries, such as transaction information for banks, this data is often of little value for accurately predicting a customer's future buying behavior or his likes and dislikes. In statistical terms, many of the inputs to CRM regression models correlate poorly with future behavior, and this problem is most acute in the financial services sector, since banks actually know very little about their customers 15 through their existing relationships.

### **Summary of the Invention**

Accordingly, the purpose of the invention is to improve the reliability of such predictions.

20 The invention provides a method of predicting future behavior of an individual by analyzing the content of internet websites already visited by that individual. By "future behavior", we mean any activity such as buying, or any action resulting from the individual's preferences, likes and dislikes.

In the context of customer relationship management exercised by a business in 25 relation to its customers, the method preferably comprises predicting customers' future behavior including their commercial requirements relating to that behavior and then communicating appropriately with selected ones of those customers.

In one preferred embodiment, with the express permission of customers, their own lists of most recently visited websites form an input to the CRM predictive models. This data is continually collected by web browsers such as Internet Explorer and Netscape: the advantage of this automation is that data collection is passive, not requiring customers to fill in tedious and lengthy questionnaires about their likes and dislikes. It is also more reliable than requiring customers to fill in such forms. A great deal can be inferred about people from their web browsing behavior, such as their interests, lifestyle and leisure activities, and this richer profile is capable of improving the predictive accuracy of CRM applications.

Thus the method preferably comprises combining text from a plurality of the visited websites, identifying a plurality of the most informative words of that text, and using data representative of those most informative words as inputs to an automated predictive model whose outputs indicate the individual's likely future behavior.

This preferably involves the step of identifying, for words of the combined text, their frequency of occurrence in the combined text and also of their occurrence in a large text corpora in the same language, and selecting as the said most informative words those whose said frequency of occurrence is significantly greater in the combined text than in the large text corpora.

Preferably, the method comprises identifying, from a database of semantic vectors derived from co-occurrence statistics, the semantic vector of each of the said most informative words, and using the semantic vectors as the said representative data.

It is preferred that the number of most informative words is predetermined so as to optimize the trade-off between a sufficient predictive accuracy and a reasonable computation time. In order to achieve such an optimum, the method can be extended to involve varying the said number of most informative words in order to determine its optimum, by re-fitting the predictive model for each value of the number and noting the predictive accuracy and the time taken. A predictive accuracy can be determined by cross-validation procedures which are well known in predictive modeling.

The invention also involves a computer program for carrying out the methods described above, and also to such a computer program stored on a data carrier, and also to data processing apparatus arranged to carry out the method described.

## 5 **Brief Description of the Drawing**

Fig. 1 is an example of the results of a cluster analysis of semantic representations of words.

## **Detailed Description**

10 In order that the invention may be better understood, examples will now be described, but it will be appreciated that the invention has many different potential applications not all of which will use all the preferred features of the examples described.

A customer relationship management system for financial services businesses will now be described. For example, it could be used by a bank offering mortgages. A customer  
15 of the bank may for example be considering buying a house, so she looks at various house buying sites on the web. The list of websites and the pages that she looks at are stored by her web browser on her home PC. Her bank has already arranged with her that they can offer her a better service if she gives them access to her web browser's store of most recently visited websites. Thus a piece of software installed on her home computer (PC) sends her browser's  
20 most recently visited websites to the bank regularly. As a consequence, the bank has several entries in her web browsing profile for the word "house" and "semi-detached" and "Lincolnshire". Vectors representing these words are used as inputs to the bank's logistic regression model which predicts who should get a mortgage offer mailshot, and it uses these highly informative pieces of information for giving this customer a high probability of  
25 needing a mortgage in the near future. The bank achieves this using its predictive model which has previously been trained using a data warehouse of past browsing behavior and mortgage buying activity. The CRM may be a simple comparison process which compares the input web behavior information against information from people who have had similar

browsing profiles in the past and have taken out a mortgage shortly afterwards. Thus this customer is included in the mailing list for the mailshot.

Thus the first step of the preferred method is to collect a file containing a list of the most recently visited websites from the customer's computer.

5        The second step is to download HTML referred to in each of the websites in the list, and to combine all the text into a single text file. Preferably, all the text is used from each site, but it would be possible to select just parts of it, such as the keywords or metatags.

      The third step is to identify the most "informative" plain text words in the HTML combined file. The degree of informativeness of a word is proportional to how much its  
10 frequency differs between its occurrences in the HTML file and in a standard large text corpora in the same language, such as the British National Corpus. Such text corpora should typically contain at least one hundred million words. The reasoning behind this is that words which occur more frequently than in normal use are likely to be significant in the context, and thus informative. The frequency of occurrence should be represented as a fraction of all  
15 words in the language corpus and the HTML file, so as to discriminate between words that occur just once in the large language corpus.

      Other methods of measuring informativeness may of course be used. The most general definition of "informativeness" would be the mutual information between the behavior being predicted and occurrences of the word in the browsed site text file. If the  
20 possible behaviors of the customer are defined as a vector of outcomes  $y_1, y_2, \dots, y_n = y$  and the frequency of word  $i$  is defined as  $x_i$ , then the mutual information between occurrence of

each word and possible behaviors is defined as  $I(x_i; y) = \sum_{y \in Y} \frac{p(x_i, y)}{p(x_i)p(y)}$  (see Cover and

Thomas, 1991, *Elements of Information Theory*, New York: Wiley). The symbol  $Y$  represents all possible values of  $y$ , i.e. all possible behaviors. In practice it would be very  
25 computationally costly to calculate  $I(x; y)$  exactly for every word  $x_i$  in the language, so faster approximations to  $I(x; y)$  have to be used, such as the keyword method defined in this specification.

The next step is to rank the words according to their informativeness and to take the top  $k$  most informative words. If the number  $k$  has already been optimized for the particular application involved, then it is regarded as a fixed number. However, the number  $k$  can be treated as a variable in order to carry out an optimizing process.

5       The next step is to look up, in a predetermined database of semantic vectors derived from co-occurrence statistics, the semantic vector for each of the top  $k$  informative words. The construction of numerical vectors that represent the "meaning" of a word, or the word's "semantic vector", is a well established technique in computational linguistics, as described in Brown, P.F., Della Pietra, V. J., de Souza, P.V. Lai, J.C. (1992). Class-based n-gram  
10   models of natural language. Computational Linguistics, 18(4), 467-479.; and also in Patel, M., Bullinaria, J. A. and Levy, J. P. (1997), Extracting Semantic Representations from Large Text Corpora, Proceedings of the Fourth Neural Computational and Psychology Workshop 1997, London; and in Christopher D. Manning, Hinrich Schutze, Foundations of Statistical Natural Language Processing, July 1999, Mit Pr. ISBN : 02621 33601.

15       The construction of the semantic vectors involves the construction of a word co-occurrence matrix that goes through a large corpus of text and counts how many times pairs of words occur together within a window of, say, 10 words. The resulting vector for each word represents the kind of verbal environment in which it occurs, and this has been shown to be a good indicator of the meaning of the words. For this reason, it is better to use the  
20   semantic vectors, as inputs to the predictive model, than the words themselves. The words alone cannot convey their meaning.

      An example of the results of a cluster analysis of the semantic representations of words is given in Fig. 1 hereto. The example is taken from Reddington, M. & Chater, N. (1997), Probabilistic and distributional approaches to language acquisition, *Trends in*  
25   *Cognitive Sciences*, 1(7) 273-289 and illustrates manually extracted low-level clusters of nouns, verbs and adverbs from a dendrogram resulting from a word level analysis of the distributional statistics of the CHILDES corpus.

In the preferred example, the semantic vectors of a large vocabulary of words in English are stored in a database, and the method involves simply looking up the semantic vector for each of the top  $k$  most informative words. The database may include vocabularies in more than one language, in which case it is necessary to select the appropriate language.

5        The  $k$  semantic vectors are appended together, and used as regressors or input variables for a single CRM predictive model. Automated predictive modeling using neural networks or statistical models or rule-based models is well known and need not be described in this specification. The logistic regression model described above is a statistical model.

Although not essential, it is a preferred feature to determine the optimum value of  $k$ .  
 10      This is carried out by increasing  $k$  from 1 upwards, iterating the steps of ranking the words according to informativeness, looking up the semantic vectors, appending the  $k$  vectors and using them as regressors. With each iteration of  $k$ , the predictive model is refitted, and the time taken to fit the model is measured; also, the predictive accuracy of the model is measured using cross-validation, a conventional technique in neural networks.

15         $k$  is optimized in the context of the particular application, trading off predictive accuracy against computational time taken.

The word co-occurrence matrix described above is obviously very large, and could be as large as  $n \times n$ , where  $n$  is the number of words in the given language. This can be reduced, to improve efficiency, by singular value decomposition, using principal components analysis  
 20      (PCA) to reduce the dimensions of the co-occurrence matrix. Reducing the dimensionality of the semantic vectors increases the speed of CRM predictive models using those vectors as inputs. Again, this is an established technique and need not be described in this specification.

Once the value of  $k$  has been optimized for a given application, it can be used as a predetermined number in future operations of the method.

25        It will be understood that the outputs of the CRM predictive model are indicative of the likely future behavior of the individual concerned. In the example of house buying and mortgage selling given above, the significant words were "house", "semi-detached" and "Lincolnshire", and the corresponding semantic vectors would be appended and fed into the

logistic statistical CRM predictive model as regressors, leading to outputs indicative of "mortgage" amongst others.

The predictive model must be set up or trained in advance. If it is a neural net, it is trained using information about real behavior resulting from previous behavior, e.g. about people (customers or otherwise) who have taken out mortgages and who previously visited websites with particular text content. If it is a statistical or a rule-based model, that information about real behavior is used to set up the model.

The web-browsing information could be just part of the input to the predictive model. Other inputs could include, for example, other customer profile information such as their age and the balances of their bank accounts.

The system is of course applicable to a wide range of customer relationship management processes. Other examples might be using web browsing behavior to indicate whether the individual takes risks or is cautious financially; and to indicate likes and dislikes in products purchased, or in types of communication, or in methods of doing business. Web browsing behavior may also indicate the number of people in the household, and possible relationships with other customers or potential customers.

It will be understood that the CRM process, including the steps identified above, would be implemented on data processing apparatus as a computer program; the computer program could be resident in a business premises, or anywhere in a network such as on the internet itself.

It will also be understood that the websites included in the list could optionally include websites not visited but linked to the visited websites. Further, it will be appreciated that information on the numbers of visits of the websites could also be used, for example to give frequently visited websites greater weight in the combined text file. If a particular website was visited three times, for example, then the text could simply be included three times in the combined HTML file. More weight could also be given to sites that have been visited recently.



**ANNEX 1**